Published in final edited form as:

J Pain. 2009 March ; 10(3): 274–281. doi:10.1016/j.jpain.2008.09.003.

Can end-of-day reports replace momentary assessment of pain and fatigue?

Joan E. Broderick, Joseph E. Schwartz, Stefan Schneider, and Arthur A. Stone

Department of Psychiatry & Behavioral Science, Stony Brook University

Abstract

This study evaluated the ability of end-of-day (EOD) ratings to accurately reflect momentary (EMA) ratings on 10 widely used pain and fatigue items. Rheumatology patients ($N=105$) completed ≥ 5 randomly scheduled EMA assessments of each item per day as well as EOD ratings. Correlations were high between EOD and EMA ratings of the five pain items ($r=.90-.92$) and somewhat lower for the five fatigue/energy items ($r=.71-.86$). To examine the ability of EOD ratings to represent a week of EMA ratings, 7 EOD ratings were averaged and correlated with EMA ($r \geq .95$ for pain items, $r=.88-.95$ for fatigue/energy items). Further, averaging only 3-5 EOD ratings achieved very high correlations with a week of EMA ratings. Within-subject correlations of EOD with mean daily EMA across 7 days confirmed patients' ability to provide daily ratings that accurately reflect their day-to-day variation in symptom levels. These EOD results were compared to traditional recall ratings collected in the same protocol. It was concluded (1) that EOD ratings were a better representation of EMA than were recall ratings, and (2) that EOD ratings across a reporting period can replace EMA for studies targeting average levels of pain or fatigue.

Perspective—This study in chronic pain patients demonstrated that end-of-day ratings of pain are highly accurate representations of average levels of pain experience across a day; Ratings of fatigue were somewhat less accurate, though still at a level that would be valid.

Keywords

pain; fatigue; momentary assessment; electronic diaries; patient reported outcomes

Introduction

There are now multiple methodologies available for assessing patient reported outcomes (PROs). The traditional method is paper questionnaires that often have reporting periods of 7 days or longer.⁸ Some years ago, Ecological Momentary Assessment (EMA) was developed to collect PROs multiple times across a reporting period.²¹ Based on both theoretical and empirical considerations, EMA has been hailed as a superior approach, because patients may have difficulty with accurately recalling the continuous stream of symptom experiences and with creating a rating that is representative of the average of those experiences.²² The PROs reported in this study are pain and fatigue/energy, two of the most common symptoms in chronic illness for which patients' self-reports are indispensable.

© 2008 The American Pain Society. Published by Elsevier Inc. All rights reserved.

Corresponding author: Joan E. Broderick, Ph.D., Department of Psychiatry and Behavioral Science, Putnam Hall, South Campus, Stony Brook University, Stony Brook, NY 11794-8790, Phone: 631-632-8083, Fax: 631-632-3165, Joan.Broderick@StonyBrook.edu.

Publisher's Disclaimer: This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final citable form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

Despite the proposed advantages of EMA, it carries a cost for researcher and patient. Momentary protocols using electronic diaries involve complex software, expensive hardware, sophisticated analytic approaches, as well as significant time on the part of the patient. Some research questions require the detail afforded by EMA, such as when the dynamic relationship between mood and pain is investigated. However, often “average” or “usual” symptom level is targeted (e.g., pain level over a week). Previous studies have shown that 7-day recall ratings for pain share approximately 50% of the variance with average EMA.^{2,166} We wondered whether end-of-day recall ratings (EOD) would be more accurate in reproducing the average of EMA ratings for the same time period. If EOD ratings are accurate summaries of EMA, they could provide a very attractive measurement alternative to the more expensive EMA method and to the less accurate recall method.

Aggregated EMA ratings captured in a random time sampling protocol are the most accurate representation of an individual’s symptom experience for that reporting period and will be treated as the “gold standard” in this study. We recognize that not everyone agrees with this. Some argue that recall ratings capture a meaningful constellation of experience and thoughts that go beyond the aggregate of momentary experience.¹⁵ However, while this is plausible, to date there is little empirical evidence to support this.

Three hypotheses are tested to determine the appropriateness of using EOD as a substitute for EMA for two reporting periods: a day and a week. First, we hypothesized that single EOD ratings would correlate highly with average EMA ratings for that day. The second hypothesis was that the average of seven EOD ratings would correlate highly with average EMA ratings for that week, thus providing a good measure of the average symptom level for the week. The third hypothesis was that within-subject correlations across days of average daily EMA and EOD ratings would be at least of moderate size, indicating the ability of EOD ratings to accurately distinguish changes in pain and fatigue/energy levels across days. If these hypotheses are supported, then it would suggest that for some applications EOD could replace EMA. A comparison of these results with previous research on the correlations of EMA with traditional 7-day recall ratings will indicate if averaged EOD ratings can provide a more accurate summary of average symptom levels than longer-term recall.

Materials and Methods

Participants

Study participants ($N=117$) were recruited from two offices of a community rheumatology practice. Patients were required to be available for 30 consecutive days and to satisfy the following eligibility criteria: ≥ 18 years of age; absence of significant sight, hearing, or writing impairment; fluent in English; normal sleep-wake schedule; diagnosed with a chronic rheumatological disease; experienced pain or fatigue in the last week; able to come to the research office two times within a month; and, had not participated in another study using an electronic diary within the last 5 years. The study protocol was approved by the Stony Brook University Institutional Review Board, and patients provided informed consent and were compensated \$100 for participation in the study. The data were collected between September 2005 and June 2006.

Telephone screening of 279 patients determined that 86 (31%) were ineligible due to visual or hearing difficulties, inability to hold a pen, atypical sleep-wake schedule, no chronic illness diagnosis, or previous participation in a momentary assessment study. Of 193 eligible patients, 76 (39%) declined participation, and 117 (61%) participated. Eleven patients dropped out of the study, and 1 patient was eliminated from the analysis due to missing data for all 7 days of the analysis, thus the study sample was $N=105$. The most prevalent diagnoses were osteoarthritis (49%), rheumatoid arthritis (29%), lupus (16%) and fibromyalgia (11%).

Participants tended to be female (87%), White (91%), married (65%), and had a mean age of 56 years (range 28-88). Most were high school graduates (96%), with 72% having completed some college (see Table 1).

Measures

Momentary and EOD ratings of pain and fatigue/energy were collected for 29-31 days on a hand-held computer (Palm Zire 31). The electronic diary (ED) utilized a software program provided by invivodata, inc. (Pittsburgh, PA) that featured auditory tones to signal the participant to complete a set of ratings. The ED was programmed to generate an average of 7 randomly-scheduled (within intervals) prompts spread across the participant's waking hours (an average of one every 2 hours and 20 minutes, constrained to ensure a minimum of 30 minutes between prompts) determined by when the participant informed the ED that she was going to bed at night and set the wake up alarm the next morning. In addition to the random signals, the ED prompted the participant to complete an EOD assessment at the time the ED was put to sleep at night.

Each ED assessment began with the participant responding to questions about his location, activity, whether alone or with others, and one positive and one negative affect rating ("happy" and "frustrated") on a visual analogue scale (VAS). Our pain and fatigue/energy items were drawn from well-established questionnaires used to assess pain and fatigue: SF36 version 2²⁶, Brief Pain Inventory (BPI)⁵, Brief Fatigue Inventory (BFI)¹², and McGill Pain Inventory^{10,11} (See Table 2). The momentary items were written in the past tense, because patients were instructed to make their rating based upon their symptom experience just before they were beeped. The EOD items were prefaced with the instruction, "please think about the entire day when answering the following questions." All momentary and EOD item ratings were made on a 100-point VAS. With the exception of the SF36v2 Bodily Pain item (*none to very severe*), all of the other item VAS anchors were *not at all* to *extremely*.

Procedure

Following a telephone eligibility interview, patients came to the research office and were trained in the use of the electronic diary (ED) to collect the momentary and EOD ratings of pain and fatigue/energy for approximately 30 days. A research assistant telephoned the patient 24 hours after the research office visit to answer any questions and troubleshoot any problems with using the ED. Additionally, a follow-up call was made once a week for the next three weeks to ensure the ED was working properly and to answer any questions. Across the month, patients also completed six recall assessments that varied by the length of the recall; these data have been reported in a previous paper.² At the end of the month, patients returned to the research office to deliver the ED and to engage in a recall assessment of the last 7 days of pain.

Analytic plan

The hypotheses were tested using the first 7 days of data in order to avoid any potential effects of extended repeated measurement on both the momentary and EOD ratings. Also, these data would best generalize to the many studies involving a reporting period of 7 days.

Construction of latent trait variables for momentary data

Analyses are based on a comparison of the average of EMA ratings with the EOD ratings for the corresponding reporting period. In order to have the most accurate representation of the reporting period, we conceptualized the mean of the EMA ratings (for a day or a week) as a latent variable, thereby adjusting for the inherent unreliability of the observed mean due to random sampling variability and missed ED prompts. Specifically, we estimated multilevel mixed models (using the MIXED procedure in SAS, version 9.1) in which the mean score of

momentary ratings for each reporting period was treated as a random (i.e., latent) variable. Had we simply used observed averages of momentary ratings in the analyses, the standard errors of the means would tend to be larger for the mean of a day than for the mean of the week due to the smaller number of momentary reports used to calculate the former. Hence, the correlation of the average momentary rating with the EOD rating would be more attenuated (i.e. weaker) for daily relative to one-week averages. The estimates generated by the multilevel modeling approach are adjusted for attenuation, and reflect our best estimate of the association between EOD measures and the “true mean” of momentary pain or fatigue/energy (while awake).

We restricted our use of latent variables to momentary assessments because they were serving as the “gold standard” for assessing the predictive utility of EOD ratings. However, when EOD assessments were aggregated to predict a 7-day average of momentary assessments, we used a simple average that did not adjust for unreliability due to sampling variability, because we had complete coverage (i.e. 7 EOD reports) for the week.

Missing data

For any given day, sufficient momentary assessments are necessary to accurately characterize symptom experience for that day. Of the 5-7 random prompts that were scheduled each day, a minimum of 3 reports was required based upon our earlier work indicating that 3 reports is comparable to a higher density of 6 or 12 reports across the day.²³ A patient’s data for any of the seven days was classified as missing if fewer than 3 momentary ratings were captured or if the EOD assessment for that day was missing. Ninety-six patients (91%) provided acceptable data for all 7 days, 7 patients had data for 6 days, and 2 had data for 5 days. All analyses were performed for the full sample of $N = 105$. For the 1-day analyses, means and correlations for each of the 7 days were simultaneously estimated in a single mixed-effects model, using the full-information maximum likelihood method. This method allows inclusion of all cases in the analysis even if some observations are missing, and has been referred to as “state of the art” for handling missing data in the methodological literature.¹⁷ For the 7-day analyses, if there were less than 7 EOD reports, ratings from all non-missing days were averaged to create a mean score for the week.

Results

Rating compliance and missing data

In order to include a patient’s data from any given day in the analyses, data were required for both the EOD and a minimum of 3 EMA ratings for that day. EMA prompts were randomly scheduled and, therefore, are viewed as a random sample of all possible reports that could have been obtained during the sampling period. Compliance was high with patients completing an average of 91% of the momentary prompts (range 68% - 100%). This compliance makes it reasonable to assume that the momentary data represent a random sampling of symptom experience over the patients’ waking hours. On average, patients completed 5.6 ($SD = 1.27$) momentary assessments each day with 82% of days having 5 or more momentary reports. Across all patients and days (735 reporting days) only 9 days had less than 3 momentary reports, and only 4 EOD assessments were missing, resulting in only 11 days (1.5%) classified as missing data (for 2 days, both the EOD assessment was missing and less than 3 momentary reports were completed).

Mean ratings on pain and fatigue items

Descriptive statistics showing the means and standard errors for each of the pain and fatigue/energy items on each of the days of the week and aggregated over the entire week are shown in Table 3. Pain intensity (BPI) as measured by EMA averaged 40.7 for the week and was significantly lower than the EOD average of 47.0 ($p < .001$). Mean momentary fatigue (BFI)

was also significantly lower than the mean EOD ratings for the week (47.0 vs. 52.5, $p < .001$). Eight of the 10 items showed this pattern for both daily and weekly average comparisons. For the two positive symptoms (energetic and full of life), the pattern was reversed, though only significantly so for the latter.

Analysis of daily correlations between EMA and EOD

The first hypothesis was that there would be a high correlation between individuals' averages of EMA ratings for a single day (treated as a latent variable) and their corresponding EOD assessments for the same day. The "average" daily correlation, pooled across the 7 days, for each of the 10 pain and fatigue/energy items are presented in Figure 1 (open diamonds). The first five items on the x-axis measure pain, and the magnitudes of their pooled daily correlations are high: pain intensity (.90), bodily pain (.90), stabbing (.91), nagging (.92), and aching (.92). The last five items measure fatigue and energy-related symptoms. These correlations are somewhat lower: fatigue (.86), tired (.83), worn-out (.82), full of life (.83), and energetic (.71). The bars around each pooled daily correlation show the range of individual correlations across the 7 days for that item. Relative to the consistently strong correlations for the pain items, a much wider range of correlations was observed for the fatigue and energy items.

Analysis of correlations between average EOD and EMA for 7 days

The second hypothesis addresses the degree of correspondence between the average of 7 days of EOD ratings with the average of the week's EMA ratings. These correlations are displayed in Figure 1 (solid diamonds). For the pain items, none was below .95 in magnitude. For fatigue/energy items, the correlations ranged from .88 to .95. These high correlations indicate that for all items the average of seven EOD assessments corresponds excellently with the average momentary symptom level for the week.

While this finding confirms the hypothesized association between EOD and EMA reports for the week, we were also interested in extending the logic of these analyses to determine if averaging fewer than 7 EOD reports can also result in a high level of correspondence with the week's average momentary symptom level. To accomplish this, we estimated the correlations that would be obtained when a single EOD report or the average of fewer than 7 (i.e., 2 through 6 randomly selected) EOD reports is used. For each EOD report that is dropped from the average, the reliability of the average decreases, which attenuates the correlation coefficient below the level found for the full set of 7 EOD reports. The reliability of the average of any specified number of EOD reports was estimated from the internal consistency of the full set of 7 EOD reports, using the Spearman-Brown formula.¹³ This reliability estimate was used, in turn, to calculate the correlation between the week's EMA average and the average of EOD reports when the latter is based on fewer than seven reports. The results are displayed in Figure 2.

For the pain items (Figure 2a), when at least 3 EOD reports are averaged, the correlation with the week's EMA average is greater than .90; averaging 5 EODs yields a correlation of about .95. For the fatigue and energy items, the strength of correspondence is lower, especially for some items (Figure 2b). The SF36v2 *energetic* item stands out as having poorer correspondence than the other items. The 4 other items reach a correlation of .90 by using 4 or 5 EOD reports, but the *energetic* item can only approach a correspondence of .90 by using all 7 EOD reports.

Analysis of within-subject correlations of EMA and EOD ratings across the 7 days

The third hypothesis addresses the ability of EOD reports to accurately capture daily variation in symptom levels across the week. For instance, if the aggregated EMA ratings were higher on 1 day compared with the remaining 6 days, was the EOD correspondingly higher on that day? These average within-subject correlations are displayed in Figure 1 (see x symbol). For

the pain items, the correlations ranged from .71 to .80, whereas for the fatigue/energy items, the correlations ranged from .46 (*energetic*) to .72.

Discussion

The objective of this study was to determine if EOD ratings of pain and fatigue are a reasonable proxy for EMA in studies interested in daily or weekly symptom levels. Our results confirmed the first hypothesis by showing that daily correlations between EOD and EMA were above .90 for the pain items and somewhat less, but still greater than .80, for most of the fatigue items. Second, the 7-day average of EOD ratings correlated very highly with the week's average EMA for all of the pain items ($\geq .95$) and almost as well for the fatigue/energy items ($\geq .88$). Furthermore, analyses suggested that fewer than seven EOD reports could be averaged and still provide a high level of correspondence with the week's EMA. Jensen and McFarland also reported the feasibility of collecting fewer ratings to accurately represent the week.⁷

The third hypothesis, a more stringent test of the validity of EOD reports, considered their ability to discriminate day-to-day fluctuations in patients' levels of pain and fatigue across the 7 days. Within-subject analyses demonstrated good correspondence for the five pain items and two of the fatigue items (r 's $\geq .70$); three of the SF36v2 items did not perform as well (r 's of .46 to .66). In comparison, we reference other within-subject correlations collected in this protocol that have been reported previously.² On the final day of the study, patients retrospectively reported for each of the last 7 days the BPI *average pain* (for the day). A within-subject analysis was conducted of those retrospective pain ratings with the daily averages of EMA ratings for those days. The average within-subject correlation was much lower ($r = .29$) for the retrospective reports compared with the average correlation of .76 for the EOD ratings of the same item. Taken together, these data indicate that patients can provide EOD ratings that reflect differing pain levels across the days, whereas retrospective reports fail to do so.

The finding that the fatigue and energy items corresponded less well with EMA than the pain items was unexpected. It is possible that fatigue and energy ratings made over some time period (hours or days) may evoke more complex considerations that go beyond the experiences captured by momentary assessment.¹⁵ For example, a patient may rate energy on an EOD report as lower if he found that he was not able to accomplish all of his tasks for the day, yet this might not be reflected in that day's momentary ratings of energy. Second, fatigue has been shown to exhibit a pronounced diurnal cycle, with greater fatigue at the end of the day, in some individuals.¹⁹ If the EOD rating is influenced by current levels of fatigue, then the inter-individual variability in diurnal cycles would be expected to attenuate its correspondence with the average fatigue for the day measured by aggregated EMA ratings.

One might still argue that less costly, single recall ratings (e.g., 7-day) are preferable to collecting EOD ratings as outcome measures. This argument can be evaluated in the context of other data collected in this same protocol and reported elsewhere² that examined the correlation of EMA with traditional recall ratings (RR) using various reporting periods (1, 3, 7, and 28-days). Those data allow us to compare the relative accuracy of EOD ratings versus RRs for the same 10 pain and fatigue/energy items. We found that the correlations of the 7-day RRs and EMA ratings for most of these items were less than .80 – and in a number of cases much less. In contrast, this study demonstrated that the average of 7 EOD ratings yielded a correlation of $\geq .95$ for the pain items and $\geq .88$ for the fatigue/energy items. Therefore, the level of shared variance between EMA and the EOD ratings ranged between 77-90%, while it ranged from 27-55% for the 7-day recall ratings.² This is a substantial difference. When a clinician or researcher wants an outcome measure that represents the patient's aggregated experience of the symptom over a given reporting period, these data provide strong support for using time- and date-stamped EOD ratings rather than recall assessment.

Evidence of the effects of item scaling on correlations with EMA was obtained by another comparison of the EOD data in this paper with the data reported in our recall paper². As noted above, one of the recall periods in the previous paper was 1-day. This 1-day RR was essentially the same as the EOD task in this study: both used the same item wording and both were made in the evening. The correlations of 1-day RR with EMA from the prior paper were generally in the range of .70 - .80, which are lower than those observed for the EOD ratings reported in this paper. In theory, they should be the same, since both involve recall over one day. What is different are the response scales used and the method for capturing the ratings (EOD=hand-held computer; RR=paper and IVR). The EOD ratings from this study were done on 100-point VAS as were the EMA ratings. Thus, the EOD and EMA ratings shared an identical scale format. In the case of the 1-day RRs from the prior study, we preserved the scaling of the item of the standard instrument. Depending upon the instrument from which the item was drawn, the 1-day RRs often had only 4-6 response options. One interpretation of these results is that the ability of RRs to finely differentiate pain and fatigue levels may be attenuated for items associated with a small number of response options. In fact, the two RR items with 0-10 point scales (BPI: average pain, and BFI: usual fatigue) showed the highest correlations with EMA ($r = .82$ and $.86$, respectively). However, a number of studies have examined VAS versus Likert scales for the measurement of pain, and there is not a consensus on whether scales with a greater number of response options perform better.^{1,9,25} Furthermore, most evidence suggests that the method (paper vs. computer) for capturing the ratings have little impact on ratings.^{3,4}

Although not specified as a hypothesis at the outset, we examined level differences in EOD and EMA reports. Consistent with previous work,^{20,24} we observed mean level differences between EOD and EMA ratings in 9 of the 10 items. While level differences between aggregated EMA and traditional RRs are frequently found, we did not know if level differences would be evident in ratings with such a proximal reporting period. The EOD ratings of the negative symptoms were generally about 5-6 points higher than the EMA ratings. This is less than the level differences observed in RRs which can be upwards of 10-15 points depending upon the item and the length of the recall period.^{2,20,27} However, it suggests that even when recall is only hours, as in EOD reports, cognitive heuristics, such as peak or duration neglect (ignoring times when the symptom is close to zero) may inflate the ratings.^{14,18} The implication of this observation is that as long as the same measurement method is used for repeated measures in a clinical trial, the level difference will remain constant and will not impact the interpretation of treatment outcomes. However, if different methods are used, there is a serious risk of false conclusions. For example, if a researcher decided to use a 1-month recall of usual pain for baseline and the average of 7 EOD ratings for post-treatment assessment, it is likely that he would observe a 10-point or more reduction in pain simply due to measurement differences. Conversely, if the EOD ratings were used at baseline and the 1-month recall at post-treatment, then pain would appear to be 10-points worse at post-treatment combined with whatever actual treatment effect was operating.

Finally, the limitations of this study are considered. First, these data may not generalize to PROs other than pain and fatigue/energy or to other populations. Second, it is possible that by virtue of providing five or more EMA symptom reports per day patients were more informed of their symptom levels and could make a more accurate EOD report. The accuracy of the EOD reports in this study may therefore be somewhat better than those that would be obtained in the absence of simultaneous EMA reports. However, it should be noted that the traditional recall reports that were compared to the EOD reports from the same protocol would have benefited in the same way. Thus, the comparative improvement in accuracy with EOD is not in doubt. Third, the patients making up this sample were rheumatology patients who on average had moderate levels of pain and fatigue. We do not believe that the patterns observed would be different in samples of pain center patients with higher symptom levels; but this would need to be confirmed empirically. Fourth, this study was not conducted in the context of a clinical

trial where symptom levels would be expected to change across the assessment period. Our patients were generally in a steady clinical state. It is uncertain as to whether EOD accuracy would be different in a clinical trial.

Acknowledgments

This research was supported by grants from the National Institutes of Health (1 U01-AR052170-01; Arthur A. Stone, principal investigator) and by GCRC Grant M01-RR10710 from the National Center for Research Resources. We would like to thank Pamela Calvanese, Doerte Junghaenel, and Leighann Litcher-Kelly for their assistance in collecting data. Software and data management services for the electronic diary assessments were provided by invivodata, inc (Pittsburgh, PA). JEB and AAS have a financial interest in invivodata, inc. and AAS is a senior scientist for the Gallup Organization.

References

1. Bolognese JA, Schnitzer TJ, Ehrich EW. Response relationship of VAS and Likert scales in osteoarthritis efficacy measurement. *Osteoarthritis Cartilage* 2003;11:499–507. [PubMed: 12814613]
2. Broderick JE, Schwartz JE, Vikingstad G, Pribbernow M, Grossman S, Stone AA. The accuracy of pain and fatigue items across different reporting periods. *Pain*. 2008 Apr 30;Epub ahead of print
3. Bushnell DM, Reilly MC, Galani C, Martin ML, Ricci JF, Patrick DL, McBurney CR. Validation of electronic data capture of the Irritable Bowel Syndrome--Quality of Life Measure, the Work Productivity and Activity Impairment Questionnaire for Irritable Bowel Syndrome and the EuroQol. *Value Health* 2006;9:98–105. [PubMed: 16626413]
4. Cook AJ, Roberts DA, Henderson MD, Van Winkle LC, Chastain DC, Hamill-Ruth RJ. Electronic pain questionnaires: a randomized, crossover comparison with paper questionnaires for chronic pain assessment. *Pain* 2004;110:310–317. [PubMed: 15275781]
5. Daut RL, Cleeland CS. The prevalence and severity of pain in cancer. *Cancer* 1982;50:1913–1918. [PubMed: 7116316]
6. Jensen MP, Mardekian J, Lakshminarayanan M, Boye ME. Validity of 24-h recall ratings of pain severity: Biasing effects of “Peak” and “End” pain. *Pain* 2008;137:422–427. [PubMed: 18035495]
7. Jensen MP, McFarland CA. Increasing the reliability and validity of pain intensity measurement in chronic pain patients. *Pain* 1993;55:195–203. [PubMed: 8309709]
8. Litcher-Kelly L, Martino S, Broderick J, Stone A. A systematic review of measures used to assess chronic pain in randomized clinical trials and controlled trials. *Journal of Pain* 2007;8:906–913. [PubMed: 17690014]
9. Lund I, Lundberg T, Sandberg L, Budh CN, Kowalski J, Svensson E. Lack of interchangeability between visual analogue and verbal rating pain scales: a cross sectional description of pain etiology groups. *BMC Medical Research Methodology* 2005;5:31. [PubMed: 16202149]
10. Melzack R. The McGill Pain Questionnaire: major properties and scoring methods. *Pain* 1975;1:277–299. [PubMed: 1235985]
11. Melzack R. The short-form McGill Pain Questionnaire. *Pain* 1987;30:191–197. [PubMed: 3670870]
12. Mendoza TR, Wang XS, Cleeland CS, Morrissey M, Johnson BA, Wendt JK, Huber SL. The rapid assessment of fatigue severity in cancer patients: use of the Brief Fatigue Inventory. *Cancer* 1999;85:1186–1196. [PubMed: 10091805]
13. Nunnally, JC.; Bernstein, IH. *Psychometric theory*. Vol. 3. New York: McGraw-Hill; 1994.
14. Redelmeier DA, Kahneman D. Patients’ memories of painful medical treatments: real-time and retrospective evaluations of two minimally invasive procedures. *Pain* 1996;66:3–8. [PubMed: 8857625]
15. Revicki DA. FDA draft guidance and health-outcomes research. *Lancet* 2007;369:540–542. [PubMed: 17307086]
16. Salovey P, Smith A, Turk DC, Jobe JB, Willis GB. The accuracy of memories for pain: not so bad most of the time. *American Pain Society Journal* 1993;2:184–191.
17. Schafer JL, Graham JW. Missing data: our view of the state of the art. *Psychological Methods* 2002;7:147–177. [PubMed: 12090408]

18. Stone A, Broderick J, Kaell A, DelesPaul P, Porter L. Does the peak-end phenomenon observed in laboratory pain studies apply to real-world pain in rheumatoid arthritis? *Journal of Pain* 2000;1:212–217. [PubMed: 14622620]
19. Stone A, Broderick J, Porter L, Kaell A. The experience of rheumatoid arthritis pain and fatigue: examining momentary reports and correlates over one week. *Arthritis Care Res* 1997;10:185–193. [PubMed: 9335630]
20. Stone A, Broderick J, Shiffman S, Schwartz J. Understanding recall of weekly pain from a momentary assessment perspective: Absolute accuracy, between- and within-person consistency, and judged change in weekly pain. *Pain* 2004;107:61–69. [PubMed: 14715390]
21. Stone A, Shiffman S. Ecological momentary assessment (EMA) in behavioral medicine. *Ann Behav Med* 1994;16:199–202.
22. Stone, A.; Shiffman, S.; Atienza, A.; Nebling, L., editors. *The science of real-time data capture: Self-reports in health research*. New York: Oxford University; 2007.
23. Stone AA, Broderick JE, Schwartz JE, Shiffman S, Litcher-Kelly L, Calvanese P. Intensive momentary reporting of pain with an electronic diary: reactivity, compliance, and patient satisfaction. *Pain* 2003;104:343–351. [PubMed: 12855344]
24. van den Brink M, Bandell-Hoekstra EN, Abu-Saad HH. The occurrence of recall bias in pediatric headache: a comparison of questionnaire and diary data. *Headache* 2001;41:11–20. [PubMed: 11168599]
25. Von Korff M, Jensen MP, Karoly P. Assessing global pain severity by self-report in clinical and health services research. *Spine* 2000;25:3140–3151. [PubMed: 11124730]
26. Ware, JE., Jr; Kosinski, M.; Turner-Bowker, D.; Gandek, B. *How to score version 2 of the SF-12 Health Survey*. Lincoln, RI: Quality Metric Incorporated; 2002.
27. Williams DA, Gendreau M, Hufford MR, Groner K, Gracely RH, Clauw DJ. Pain assessment in patients with fibromyalgia syndrome: a consideration of methods for clinical trials. *Clin J Pain* 2004;20:348–356. [PubMed: 15322442]

Abbreviations

BFI	Brief Fatigue Inventory
BPI	Brief Pain Inventory
ED	electronic diary
EMA	ecological momentary assessment
EOD	end of day assessment
IVR	Interactive Voice Recording
PRO	patient reported outcome
RR	recall rating across a reporting period

Figure 1.

The black diamond is the correlation of the average of 7 EOD ratings with the average of the week's EMA ratings. The open diamond represents the pooled average of the 7 correlations between EOD ratings and corresponding mean EMA ratings for that day, the error bars denote the range in correlations. The x is the pooled within-subject correlation of EOD and daily averages of EMA reports across the 7 days. EMA averages are treated as latent variables.

Figure 2.

Estimated correlations of one-week average momentary symptom levels (mean EMA, treated as a latent variable) with the average of EOD reports, as a function of the number of EOD reports (1 to 7) being averaged. Correlations are based on the Spearman-Brown formula, with correction for finite sampling (maximum of 7 EOD day reports in a week).

Table 1

Demographic characteristics of the sample (N=105)

Age (Mean, SD)	55.6 (11.1)
Female	87%
Race (White)	91%
Married	65%
Education	
Did not finish high school	4%
High school graduate	24%
Some college	30%
College graduate or more	42%
Employed	51%
Disability benefits	25%
Family income	
Less than \$20,000	10%
\$20,000 - \$49,999	26%
\$50,000 - \$74,999	28%
Greater than \$75,000	35%

Table 2
Random Momentary & EOD Items Mapped Onto Original Questionnaire Items

Source of item	Original Questionnaire Item	Momentary Items	EOD Items	100-point VAS anchors
Pain Items				
SF36v2	How much <u>bodily</u> pain have you had during the past 4 weeks?	How much bodily pain did you have?	How much bodily pain did you have?	none, very severe
BPI	Please rate your pain by circling the one number that best describes your pain on the <i>average</i> .	How intense was your bodily pain?	What was the average level of your pain today?	not at all, extremely
McGill BPI	Aching (sensory factor)	How aching was your pain?	How aching was your pain?	not at all, extremely
McGill BPI	Stabbing (sensory factor)	How stabbing was your pain?	How stabbing was your pain?	not at all, extremely
McGill BPI	Nagging (affective factor)	How nagging was your pain?	How nagging was your pain?	not at all, extremely
Fatigue Items				
SF36v2	How much of the time during the past 4 weeks did you feel full of life?	How full of life did you feel?	How full of life did you feel?	not at all, extremely
SF36v2	How much of the time during the past 4 weeks did you have a lot of energy?	How energetic did you feel?	How energetic did you feel?	not at all, extremely
SF36v2	How much of the time during the past 4 weeks did you feel worn out?	How worn out did you feel?	How worn out did you feel?	not at all, extremely
SF36v2	How much of the time during the past 4 weeks did you feel tired?	How tired did you feel?	How tired did you feel?	not at all, extremely
BFI	Please rate your fatigue (weariness, tiredness) by circling the one number that best describes your USUAL level of fatigue during the past 24 hours.	How fatigued (weary, tired) did you feel?	What was the usual level of your fatigue today?	not at all, extremely

Note. SF36 version 2 (SF36v2), Brief Fatigue Inventory (BFI), Brief Pain Inventory (BPI), McGill Pain Questionnaire (McGill).

Table 3
Means (standard errors) of Pain and Fatigue Measures for each of 7 Days and for the Week

	Days														Week	
	Day 1		Day 2		Day 3		Day 4		Day 5		Day 6		Day 7			
	EMA	EOD	EMA	EOD	EMA	EOD	EMA	EOD	EMA	EOD	EMA	EOD	EMA	EOD	χ^2a	$t_{1, 103}$
Pain Intensity	39.6 (2.3)	45.7 (2.6)	41.4 (2.2)	48.6 (2.6)	38.9 (2.3)	44.4 (2.4)	40.3 (2.2)	44.7 (2.5)	42.0 (2.2)	47.8 (2.3)	41.6 (2.2)	49.2 (2.4)	41.4 (2.2)	48.4 (2.6)	180.7***	10.3***
Body Pain	43.0 (2.0)	48.0 (2.3)	44.5 (2.0)	51.4 (2.5)	42.0 (2.1)	47.8 (2.4)	43.3 (2.0)	47.0 (2.3)	45.6 (2.0)	51.1 (2.2)	43.7 (2.0)	49.6 (2.2)	44.1 (2.1)	51.2 (2.2)	172.8***	9.1***
Stabbing	22.3 (2.0)	25.2 (2.6)	24.9 (2.4)	28.0 (2.9)	22.2 (2.2)	26.5 (2.7)	23.1 (2.2)	26.4 (2.6)	23.7 (2.4)	25.4 (2.8)	22.7 (2.4)	24.6 (2.6)	21.2 (2.3)	23.8 (2.6)	39.6***	4.9***
Nagging	39.7 (2.3)	43.2 (2.9)	40.2 (2.4)	44.5 (3.0)	38.6 (2.5)	42.6 (2.7)	39.6 (2.5)	42.8 (2.7)	40.9 (2.6)	45.6 (3.0)	40.1 (2.6)	45.1 (2.9)	39.9 (2.6)	46.6 (3.0)	92.4***	6.6***
Aching	43.1 (2.2)	45.4 (2.5)	44.7 (2.2)	50.9 (2.6)	41.8 (2.2)	48.1 (2.6)	43.0 (2.3)	47.1 (2.5)	45.2 (2.3)	50.5 (2.4)	44.1 (2.3)	49.6 (2.5)	45.2 (2.4)	51.4 (2.6)	146.4***	7.8***
Fatigue	47.5 (2.0)	55.5 (2.7)	48.1 (2.1)	55.2 (2.7)	45.7 (2.1)	52.3 (2.5)	46.0 (2.1)	52.8 (2.5)	47.4 (2.1)	49.9 (2.7)	48.6 (2.0)	51.3 (2.6)	45.8 (2.1)	50.3 (2.5)	96.4***	7.3***
Tired	49.8 (1.9)	58.4 (2.4)	48.2 (2.1)	56.6 (2.5)	45.5 (2.0)	51.1 (2.5)	47.1 (1.8)	52.7 (2.3)	47.0 (2.0)	52.1 (2.4)	47.5 (2.0)	53.4 (2.4)	45.9 (2.0)	54.5 (2.4)	147.3***	8.4***
Worn out	45.5 (1.9)	54.4 (2.6)	46.0 (2.0)	53.0 (2.5)	44.5 (2.0)	52.2 (2.5)	44.9 (2.0)	50.3 (2.5)	45.7 (2.0)	48.1 (2.6)	47.0 (2.0)	50.4 (2.6)	44.0 (2.0)	50.9 (2.5)	101.9***	8.4***
Energetic	39.1 (1.5)	38.9 (2.1)	41.5 (1.7)	40.5 (2.1)	42.2 (1.6)	39.9 (2.2)	41.6 (1.5)	41.4 (2.2)	41.4 (1.5)	40.8 (2.3)	40.5 (1.7)	39.5 (2.0)	41.6 (1.7)	40.2 (2.1)	4.0	-1.3
Full of Life	47.4 (1.8)	47.2 (2.2)	49.3 (1.9)	48.3 (2.2)	49.4 (1.9)	47.6 (2.3)	49.5 (2.0)	45.7 (2.4)	48.4 (1.8)	47.6 (2.2)	47.7 (1.8)	45.5 (2.2)	47.7 (1.9)	45.0 (2.3)	14.9*	-2.8**

Note: Nparticipants = 105. EMA = Average ecological momentary assessment (latent variable). EOD = End of day assessment. 7EOD = Observed average of 7 end of day assessments.

a Wald test statistic for the multivariate difference between EMA and EOD ratings across days ($df = 7$).

* $p < .05$;
** $p < .01$;
*** $p < .001$